

---

# Appendix D

## TRI Data Quality Program

---

The goals of EPA's data quality program for TRI are to: (1) identify and assist facilities that must report so that data submitted will be of the highest quality, (2) ensure high quality data entry, (3) correct and normalize as much of the submitted data as possible in order to maximize the utility of the data, (4) accurately assess the relative validity of release estimates and other data, and (5) ensure completeness of the database with compliance and enforcement measures.

### IDENTIFICATION AND ASSISTANCE TO FACILITIES

Through work with a wide variety of trade associations, local and national seminars, training courses, and enforcement activities, EPA has endeavored to locate all facilities required to report under section 313 of EPCRA and inform them of their obligations. In addition, EPA has prepared various materials to assist facilities in complying with EPCRA section 313. These include detailed reporting instructions, a question-and-answer document, magnetic media reporting instructions, general technical guidance, and 27 industry-specific guidance documents. In addition, EPA maintains a toll-free hotline to answer regulatory and technical questions to assist facilities.

### DATA ENTRY QUALITY ACTIVITIES

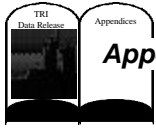
EPA continues to place a high emphasis on data entry accuracy within the Toxics Release Inventory database. For the 1997 reporting year, EPA's internal review of approximately 4% of the records showed a data entry accuracy rate of over 99.9%. This is up from a 1987 reporting year rate of 97.5%. EPA

continued the computerized edit checks at the point of data entry, including a high percent of verification and formalization of data reconciliation activities. EPA mailed copies of the release and transfer estimates to all reporting facilities to allow them to verify the entered data. EPA also received 68% of the 1997 submissions from facilities reporting on magnetic media, which ensures that fewer EPA data entry errors occur. This compares to 13% magnetic media submissions for reporting year 1990 – the second year that magnetic media reporting was available. EPA is continuing to encourage the use of magnetic media by all submitters.

### CORRECTIONS AND NORMALIZATIONS OF DATA

Because Congress has required that EPA make the TRI data available to the public through computer telecommunications, EPA has found it necessary to undertake a variety of activities to make the data more usable. This is due to the fact that computers only retrieve data in exactly the format requested (e.g., if asked for "Los Angeles," the computer will not identify facilities listed under "LA"), and facilities report their data in a wide variety of ways. As a result, EPA has taken steps to use a consistent name for all counties, has used a variety of nomenclature standards for names within the database, and has taken other steps to assist in the utilization of the data.

EPA generates a facility identification number at the time of data entry. Linkage between all years of reports has been made to the best of EPA's ability. This allows easy retrieval of cross-year data, even when a facility is sold or changes its name. The identification number has been sent to all reporting



## Appendix D – TRI Data Quality Program

facilities. Facilities are required to use this number on all future TRI reports submitted to the Agency. Use of this number facilitates data quality and cross-year analysis.

In 1998, EPA provided all states with a listing of facilities that reported for 1997 to verify that both the state and federal government received the same data. States that responded found cases where facilities had not reported to one or the other government. States provided copies of forms to the EPA where EPA had not received copies, and vice-versa. This activity has provided a critical step to assist EPA in coordinating the data collection with the states.

Every year EPA issues Notices of Noncompliance (NONs) to facilities who use invalid forms or provide incomplete forms, incomplete facility identification, or incorrect/missing chemical identification. These facilities are also notified by telephone to make sure their follow-up revisions correct these errors. A facility that does not comply with a NON may be subject to civil penalties.

For reporting year 1997, EPA has again issued Notices of Technical Error (NOTEs) for missing required data or for incorrect information, such as facility identification numbers or invalid codes. The response rate to the NONs and NOTES has been very good and has prevented errors from recurring in following years. In addition, to help facilities avoid these types of errors, EPA provides a document entitled *Data Quality Checks to Prevent Common Reporting Errors on Form R/Form A*.

### ACCURACY EVALUATION

TRI data are widely used by the public, the media, state, local and tribal governments, environmental and industry advocacy groups, researchers, and the business community. The number of TRI data user groups grow every year. Therefore, TRI data quality is an important issue for any meaningful data analysis. The Agency is currently using data validation

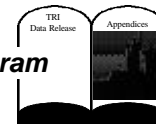
techniques, protocol to approve or deny any request to withdraw TRI data from the public database, TRI voluntary site surveys to improve TRI data quality, and TRI inspections. In order to ensure high quality, the Agency is in the process of setting up a procedure, similar to all withdrawal requests, to revise the data submitted to TRI.

The accuracy of the release data can vary. Some releases can be estimated fairly easily, just by knowing how much of a chemical was used or by weighing drums of solid/liquid waste, for example. Where monitoring of release streams or wastes has been done, release estimates may be within 20% of the actual amount released, although infrequent, non-representative sampling may lead to much less accuracy. Estimates of fugitive air emissions and complex wastewaters for which monitoring data are not available may be off by one or even two orders of magnitude, particularly when the release is a small percentage of the amount of the chemical actually processed.

The purpose of a data quality site survey is to assess the quality of the data collected under section 313 and to identify areas where improved guidance would be useful for improving the accuracy of future reported data. Site surveys are also designed to identify the frequency and the magnitude of errors in the Form R data and the reasons these errors occurred. EPA also conducts train-the-trainer courses for industry to assist them in reporting to TRI.

For the 1987 and 1988 reporting years, EPA conducted data quality site surveys at facilities to determine how well facilities complied with the law and estimated release quantities. These surveys did not “confirm” estimates through monitoring, but determined how well facilities used available data and estimation techniques to calculate releases.

Overall, based on the survey of 156 facilities, 1987 total annual releases appeared to have been underestimated by 2%, representing the net effect of



overestimates and underestimates. For non-zero release estimates, more than three-quarters were within a factor of two of EPA's best estimate. About 15% were in error by an order of magnitude or more.

The survey of the 1988 data focused on facilities in Standard Industrial Classification (SIC) codes 28 (chemical manufacturing), 29 (petroleum refining), and 34 (metal finishing and fabrication). Ninety facilities were visited. The aggregate 1988 release estimates in these industries were more accurate than their 1987 estimates, since their aggregate 1988 estimates were found to be approximately equal to the estimates calculated by the EPA contractor.

For the 1987 and 1988 reporting year, in a different type of survey, EPA also identified approximately 1,800 forms with suspect release data and telephoned facilities to discuss how to improve and correct their estimates. The information from this survey was also used to improve the reporting instructions and technical guidance.

EPA completed a 1994 and 1995 data quality site survey for facilities in SIC codes 25 (furniture manufacturing), 26 (pulp and paper manufacturing), 28 (chemical manufacturing), and 30 (rubber and plastics manufacturing). Overall, the survey found a

high degree of agreement between facility and surveyor estimates. General trends noted in the RY 1994 data were that the total releases claimed by the facility for all SIC codes surveyed were less than the total releases claimed by the site surveyors. RY 1995 data showed that the total releases and other waste management quantities claimed by the facility approximately equaled those quantities claimed by the site surveyors. Total aggregate releases and other waste management quantities calculated by facilities and site surveyors for all SIC codes surveyed in RY 1994 and RY 1995 differed by 4%.

EPA recently completed a 1996 data quality site survey for facilities in SIC codes 33 (primary metals industry), 36 (electronic and other electrical equipment industry), and 37 (transportation equipment industry). Facilities determined thresholds correctly 95% of the time for all TRI chemicals used on-site. Overall, facilities correctly identified release and other waste management activity quantities. The survey identified facilities in the primary metals industry and electrical equipment industry (SIC codes 33 and 36) that misreported release and other waste management quantities of TRI chemicals, primarily due to the confusion over definitions of recycling.